#### **Project Overview**

**Objective:**

* This capstone project is designed to reinforce your understanding of the core components of an end-to-end data pipeline. The focus is on conceptualizing and designing a comprehensive data pipeline that addresses data ingestion, processing, storage, and analysis. The project challenges you to apply theoretical knowledge to a practical scenario without the need to build or code the actual pipeline.

**Duration:**

* 1 week (To be submitted on Saturday 24th of August, 2024)

**Deliverables:**

* A presentation deck documenting the design and architecture of the data pipeline, including the rationale behind key decisions

**Mode of Submission:**

* A link to a presentation deck to be submitted using this form: <https://forms.gle/51RSYaCDHBFDAkDk8>

#### 

#### **Project Scenario**

**Background:**

* You have been hired by a healthcare organization, "**CoreHealthCare**," which is looking to build a robust data pipeline to handle their increasing volume of patient data, clinical records, and operational metrics. The organization wants to streamline data ingestion, ensure data quality, optimize data processing, and enable real-time analytics to support decision-making.

**Current State:**

* CoreHealthCare collects data from various sources, including electronic health records (EHRs), IoT devices (wearables, medical equipment), and third-party APIs (insurance, public health databases).
* Data is currently stored in siloed databases, leading to inefficiencies in data access and analysis.
* The organization faces challenges in processing data in real-time and integrating diverse data formats.
* CoreHealthCare is exploring cloud-based solutions to enhance scalability and performance.

#### 

#### 

#### **Project Tasks**

##### **Task 1: Data Ingestion Strategy**

**Objective:**

* Design a strategy for ingesting data from multiple sources into a central repository.

**Requirements:**

* Identify and classify data sources (e.g., structured) and determine the appropriate ingestion methods for each (e.g., batch processing).
* Propose tools and technologies for data ingestion.
* Address any data consistency, latency, and reliability concerns during ingestion.
* What are the important questions you will ask to ensure a robust ingestion process.

**Deliverable:**

* A detailed data ingestion plan, including diagrams of the data flow and justifications for chosen methods and tools.

##### **Task 2: Data Processing and Transformation**

**Objective:**

* Design the data processing layer to clean, transform, and enrich the ingested data.

**Requirements:**

* Define the data processing pipeline stages (e.g data cleaning) and what is done at each stage.
* Consider both batch and real-time processing needs and select appropriate processing frameworks or technology (e.g., Apache Spark for batch processing)
* Address challenges related to data quality, such as validating incoming data.
* Ensure the design supports scalability and can handle varying data volumes and processing loads.

**Deliverable:**

* A processing pipeline design, with detailed explanations of each processing stage, the technologies used, and how the pipeline ensures data quality and efficiency.

##### **Task 3: Data Storage and Management**

**Objective:**

* Design the data storage architecture that supports efficient querying, analysis, and long-term data management.

**Requirements:**

* Choose appropriate data storage solutions for different types of data (e.g., data lake for raw data)
* Consider techniques to optimize storage performance and cost.
* Plan for data lifecycle management (e.g data retention policies)
* Address security and access control mechanisms to protect sensitive data, ensuring compliance with relevant regulations.

**Deliverable:**

* A comprehensive data storage architecture plan, including storage technologies, data organization strategies, and security measures.

##### **Task 4: Pipeline Orchestration and Monitoring**

**Objective:**

* Design the orchestration and monitoring framework for managing and maintaining the data pipeline.

**Requirements:**

* Propose an orchestration strategy to automate and manage the data pipeline workflows.
* Design a monitoring strategy to track pipeline performance, detect failures, and trigger alerts or automated recovery processes.
* Consider logging mechanisms for auditing and troubleshooting pipeline issues.
* Ensure that the design includes capabilities for scaling the pipeline as data volumes grow.

**Deliverable:**

* A detailed plan for pipeline orchestration and monitoring, including the tools, processes, and strategies to ensure a reliable and efficient data pipeline.

##### 

##### 

##### **Task 5: Presentation**

**Objective:**

* Present your end-to-end data pipeline design to CoreHealthCare’s leadership team.

**Requirements:**

* Create a presentation that summarizes the design of each component of the pipeline, from ingestion to orchestration.
* Explain the rationale behind your design choices and how they address CoreHealthCare’s challenges.
* Discuss how your pipeline design ensures data quality, scalability, and performance.
* Talk about any potential challenges and how they would be mitigated.

**Deliverable:**

* A slide deck

### **Conclusion**

This task challenges you to apply your knowledge of data engineering fundamentals by designing an end-to-end data pipeline for a healthcare organization. By focusing on each core component of the pipeline—data ingestion, processing, storages, and orchestration—you will reinforce your understanding of how these elements work together to create a cohesive data engineering solution.